A Nearly Linear–Time General Algorithm for Bi–Allele Haplotype Phasing

Will Casey[1] and Bud Mishra[1,2]

[1] Courant Institute, New York University,
251 Mercer Street,
New York, NY 10012, USA,
{gill,mishra}@cs.nyu.edu,
WWW home page: http://www.cs.nyu.edu/cs/faculty/mishra/
[2] Watson School of Biological Sciences, Cold Spring Harbor Laboratory,
Demerec Building, 1 Bungtown Road,
Cold Spring Harbor, NY 11724, USA

Abstract. The determination of feature maps, such as STSs[3], SNPs[4] or RFLP[5] maps, for each chromosome copy or *haplotype* in an individual has important potential applications to genetics, clinical biology and association studies. We consider the problem of reconstructing two haplotypes of a diploid individual from genotype data generated by mapping experiments, and present an algorithm to recover haplotypes. The problem of optimizing existing methods of SNP phasing with a population of diploid genotypes has been investigated in [7] and found to be NP-hard. In contrast, using single molecule methods, we show that although haplotypes are not known and data are further confounded by the mapping error model, reasonable assumptions on the mapping process allow us to recover the co-associations of allele types across consecutive loci and estimate the haplotypes with an efficient algorithm. The haplotype reconstruction algorithm requires two stages: Stage I is the detection of polymorphic marker types, this is done by modifying an EM–algorithm for Gaussian mixture models and an example is given for RFLP sizing. Stage II focuses on the problem of *phasing* and presents a method of local maximum likelihood for the inference of haplotypes in an individual. The algorithm presented is nearly linear in the number of polymorphic loci. The algorithm results, run on simulated RFLP sizing data, are encouraging, and suggest that the method will prove practical for haplotype phasing.

1 Introduction

Diploid organisms carry two mostly similar copies of each chromosome, referred to as *Haplotypes*. Variations in a large population of haplotypes at specific loci

[3] sequence tag sites
[4] single nucleotide polymorphisms
[5] restriction fragment length polymorphisms

are called *Polymorphisms*. The co-associations of these variations across the loci indices are of intense interest in disease research. Genetic markers such as RFLPs[6] and SNPs [7] are the units to which this paper's association studies apply.

The problems and difficulties of inferring diploid haplotypes through the use of population data have been extensively investigated ([4], [6], [10], [13], [7]), and widely acknowledged.

Our approach focuses on the use of multiple independent mapping experiments (on, for example, a collection of large DNA fragments) as the base data to infer haplotypes. Single molecule methods and technologies, such as optical mapping and polony ([9]), may accommodate the high-through-put for haplotyping diploids in a population.

We consider the problem of reconstructing two haplotypes from genotype data generated by general mapping techniques, with a focus on single molecule methods. The genotype data is a set of observations $D = \langle d_i \rangle_{i \in [1 \ldots N]}$. Each observation is derived from one of the two distinct but unknown haplotypes. Each observation $d_i = \langle d_{ij} \rangle_{j \in [1 \ldots M]}$ is a set of observations over the loci index j with $d_{ij} \in \mathbb{R}^r$.

Mapping processes are subject to noise and we assume a Gaussian model $d_{ij} \sim N(\mu, \sigma)$ with parameter μ depending on the underlying haplotype of d_i. Mapping processes shall be designed to discriminate the polymorphic allele types in the data space for each loci; hence the set of observation points $\langle d_{ij} \rangle_{i \in [1 \ldots N]}$ are derived from a mixed distribution which displays bi-modal characteristics in the presence of a polymorphic feature. By estimating the parameters of the distribution, we can assign a posteriori distribution that a particular point in $\mathbb{R}^r$ is derived from an allele type.

Since the mapping errors for d_{ij} and $d_{ij'}$ are assumed to be independent, computing the posteriori distribution for haplotypes with product allele types is straightforward, and is a major advantage of utilizing single molecule methods in association studies.

The *Phasing* problem is to determine which haplotypes are most likely generating the observed genotype data. The challenge is to infer the most likely parameter correlations across the loci index accounting for the posteriori.

2 Mapping Techniques

We wish to present applications for a wide spectrum of mapping techniques, to allow a large number of polymorphic markers (SNPs, RFLPs, micro-insertions and deletions, microsatilite copy numbers) to be used in an association study.

This paper focuses on mapping techniques capable of 1) discriminating alleles at polymorphic loci and 2) providing haplotype data at multiple loci. A mapping technique designed for association studies should be discriminating: for each polymorphic loci, data points in the data space $\mathbb{R}^r$ which are derived from separate allele types should form distinct clusters in the data. A technique which

[6] Restriction Fragment Length Polymorphisms.
[7] Single Nucleotide Polymorphisms.

allows observation of a single haplotype over multiple loci may be necessary for an efficient phasing algorithm. Thus single molecule methods are of particular interest to us. Our models and analysis are influenced by their applicability to association studies.

As an example, consider the length between two restriction fragments. The observable x is modeled as a random variable depending on the actual distance μ.

$$P(x|\mu) = \frac{1}{\sqrt{2\pi\mu}} \exp\left(\frac{-(x-\mu)^2}{2\mu}\right)$$

Isolate a specific pair of restriction sites on one of the haplotypes H_1, and let the distance between them be given by μ_1. The distance between the homologous pair on the second haplotype H_2 is given by μ_2. An observation x from the genotype data is then either derived from H_1 or H_2, denoted $x \sim H_1$ and $x \sim H_2$ respectively.

$$\begin{aligned} P(x) &= P(x|x \sim H_1)P(x \sim H_1) + P(x|x \sim H_2)P(x \sim H_2) \\ &= \frac{1}{\sqrt{2\pi\mu_1}} \exp\left(\frac{-(x-\mu_1)^2}{2\mu_1}\right) P(x \sim H_1) \\ &\quad + \frac{1}{\sqrt{2\pi\mu_2}} \exp\left(\frac{-(x-\mu_2)^2}{2\mu_2}\right) P(x \sim H_2). \end{aligned}$$

With the RFLP sizing mapping technique, observable $d_{ij}, d_{ij'}$ have independent error sources depending on loci-specific parameters. The set $\{d_{ij}, i \in [1 \ldots N]\}$ provides points in $\mathbb{R}$ which may be discriminated using a Gaussian Mixture model. Due to the uncertainty of mapping and underlying haplotypes we have chosen to model data as posteriori distribution $\alpha(x) = [P(x \sim H_1), P(x \sim H_2)]$ rather than determined allele types.

This paper is organized into five sections: section 1 defines the problems we are addressing; section 2 explains the EM-Algorithm application; section 3 discusses the phasing problem; section 4 outlines algorithm implementation and provides examples; and section 5 briefly discusses results, technologies and future work.

2.1 EM–Algorithm for detection of bi–allelic polymorphisms

The use of the EM-Algorithm for inferring parameters of a Gaussian mixture model is a well-known method (see [5] [12]), and useful in this context as well. We postulate that in the presence of polymorphisms at loci j, informative mapping data will display a bi-modal distribution in the data space $\mathbb{R}^r$. Detailed computations for the E-Step and M-Step are provided in the appendix. For each locus j the EM-algorithm is run until convergence occurs; the result being: $\langle \alpha_k(x), \hat{\Phi} = \langle \hat{\mu}_1, \mu_2, \ldots, \mu_K, \sigma \rangle \rangle$. Here α is a posteriori probability that data point x is derived from allele type $k \in [1, 2, .., K]$. **Criteria for Polymorphisms**

Let $\hat{\Phi}(D)$ denote the limit of the EM–algorithm with data set D at the loci j. A critical question is: When will a loci exhibit 2 specific allele variations? Setting $K = 2$ for the remainder of the paper (hence $\hat{\Phi} = \langle \mu_1, \mu_2, \sigma \rangle$, we define polymorphic loci as events:

$$X(D) = \begin{cases} 1 \text{ if } \hat{\Phi}(D) : |\hat{\mu}_1 - \hat{\mu}_2| - \delta > 0 \\ 0 \text{ o.w.} \end{cases}$$

Robustness of the EM-algorithm Mapping techniques contain errors that are Gaussian across a diverse set of technologies. Genetic markers may be associated or linked to allele types in the population. The mixture model treated with EM works well, sometimes distinguishing fits beyond visual accuracy. The constraint for a single value of σ forces the EM results toward one of two steady states, $\mu_1 \neq \mu_2$ or $\mu_1 = \mu_2$ (a single Gaussian). Although the EM estimates are slightly biased, the estimators are consistent and the bias is known to diminish with larger data sets.

The individual experiment data $\{d_{ij} : i \in [1...N]\}$ is mapped to posteriori probability measures over the allele classes producing a probability function $\alpha(y)$ reflecting our confidence (in the presence of mapping error) that point y corresponds to one of our allele types. For polymorphims assignments, false positives are unlikely to disturb the phasing, while false negatives affect the size of phased contigs.

3 Phasing Genotype Data

Phasing is the problem of determining co–association of alleles, due to linkage on the same haplotype. Letting Λ_j be the allele space at loci j, a haplotype may be considered an element of the set: $\prod_{j \in [1,2,...,M]} \Lambda_j$.

In phasing polymorphic alleles for an individual's genotype data (a mix of two haplotypes), we assume that half of the data is derived from each of the underlying haplotypes H_1 and H_2. In this context haplotypes have a complementary structure in that the individual's genotype must be hetero-zygote at each polymorphic loci.

We define a haplotype space and discuss how to estimate the probability that an observation d_i is derived from a particular haplotype over a set of loci. Finally, we formulate the *maximum likelihood* problem for haplotype inference, this being our solution to the phasing problem.

3.1 Haplotype Space and Joint Distributions

The full space of haplotypes is the product over all allele spaces $\{1, 2, \ldots, M\}$; in the problem under discussion the haplotype space is in one-to-one correspondence with $\mathcal{M} = \{-1, 1\}^M$. The discrete-measure space $\langle \mathcal{M}, 2^{\mathcal{M}} \rangle$ will be used to denote the haplotypes, while $\mathcal{M}[j_1, j_2, ..., j_v]$ denotes the haplotypes over the

range of loci $j_1, j_2, ..., j_v$. The result of phasing genotype data is a probability measure on the space $\langle \mathcal{M}, 2^{\mathcal{M}} \rangle$. Noiseless data may result in a measure assigning $\frac{1}{2}$ to each of the complementary haplotypes, and 0 to all others. This uniform measure over complements corresponds to perfect knowledge of what the haplotypes are. Our algorithm is consistent in that the correct result is achieved for suitably large data sets.

Let Λ_j be the allele set for the polymorphic loci j. Consider two bi-allelic loci j and j'. For clarity, we will assume that $\Lambda_j = \{A, a\}$ while $\Lambda_{j'} = \{B, b\}$. A data observation d_i is derived from one of the four classes: AB, Ab, aB, ab. Because the mapping noise at loci j and j' are independent, we can assess the probability based on the loci posteriori that the observation is derived from the following four classes:

$$
\begin{aligned}
&P(d_i \sim AB) = \alpha_{jA}(d_{ij})\alpha_{j'B}(d_{ij'}) \\
&P(d_i \sim Ab) = \alpha_{jA}(d_{ij})\alpha_{j'b}(d_{ij'}) = \alpha_{jA}(d_{ij})(1 - \alpha_{j'B}(d_{ij'})) \\
&P(d_i \sim aB) = \alpha_{ja}(d_{ij})\alpha_{j'B}(d_{ij'}) = (1 - \alpha_{jA}(d_{ij}))\alpha_{j'B}(d_{ij'}) \\
&P(d_i \sim ab) = \alpha_{ja}(d_{ij})\alpha_{j'b}(d_{ij'}) = (1 - \alpha_{jA}(d_{ij}))(1 - \alpha_{j'B}(d_{ij'}))
\end{aligned}
$$

We define $\alpha^{(i)}_{jj'}$ as the *estimated probability distribution for observation i* on haplotypes over the loci j, j':

$$\alpha^{(i)}_{jj'} = [\alpha_{jj'AB}(d_i), \alpha_{jj'Ab}(d_i), \alpha_{jj'aB}(d_i), \alpha_{jj'ab}(d_i)]$$

We define $\alpha_{jj'}$ as the *estimated probability distribution over the data set* on haplotypes over the loci j, j':

$$\alpha_{jj'}(D) = \frac{1}{N} \sum_{i=1}^{N} \alpha^{(i)}_{jj'}$$

For $\rho \in \mathcal{M}[j_1, j_2, ..., j_M]$ and $\alpha_{j_w \rho_w}(d_i) = \text{ Prob}(d_i \sim \rho_w)$ with $\rho_w \in \Lambda_{j_w}$, we can extend the estimates to any set of indices producing:

$$
\begin{aligned}
\alpha^{(i)}_{j_1 j_2 ... j_v} &= \left[\prod_{w \in [1...v]} \alpha_{j_w \rho_w}(d_i) \right]_{\rho \in \mathcal{M}[j_1, j_2, ..., j_v]} \\
\alpha_{j_1 j_2 ... j_v} &= \frac{1}{N} \sum_i \alpha^{(i)}_{j_1 j_2 ... j_v}
\end{aligned}
$$

3.2 Complementarity

In phasing the diploid genotype data into two haplotypes $\rho_1, \rho_2 \in \mathcal{M}$ there is a special property: haplotype ρ_2 is complementary to haplotype ρ_1, denoted $\bar{\rho}_2 = \rho_1$. The complementary pair of haplotypes may be represented by a change of variables, $w \in \{-1, 1\}^{M-1}$, and the transformation to the haplotypes is given by the map:

$$\rho_1(b) = \begin{cases} -1 \text{ if } b = 1 \\ -1 \prod_{j=1:(b-1)} w(j) \text{ for } b \in [2...M] \end{cases}$$
$$\rho_2(b) = \begin{cases} 1 \text{ if } b = 1 \\ 1 \prod_{j=1:(b-1)} w(j) \text{ for } b \in [2...M] \end{cases}$$

In evaluating the data, there are a possible 2^{M-1} complementary pairs of allele types to search.

The confidence of a set of complementary haplotypes is modeled as a probability distribution on the discrete measure space $\langle \mathcal{M}, 2^{\mathcal{M}} \rangle$, which is the convex hull of the following set of extremal points which correspond to certain knowledge of complementary haplotypes.

$$A = \left\{ \theta_\rho : \theta_\rho(\delta) = \begin{cases} \frac{1}{2} \text{ if } \delta = \rho \\ \frac{1}{2} \text{ if } \delta = \bar{\rho} \\ 0 \text{ o.w.} \end{cases} \quad \text{for } \delta \in \mathcal{M} \right\}$$

These values are the uniform distribution over complementary haplotypes and geometrically are vertices of a high dimensional hyper-cube. Let $A[j_1, j_2, ..., j_v]$ be the corresponding distribution over the haplotype space $\mathcal{M}[j_1, j_2, ..., j_v]$.

3.3 Maximum Likelihood Problem

We assume that for every loci j, the data $\{d_{ij} : i \in [1 \dots N]\}$ contains an equal distribution of data from the underlying haplotypes H_1, H_2 that can be inferred. Using the estimated values α for the joint distribution over loci product spaces, we will compute the haplotypes most likely producing α. We formulate the corresponding maximum likelihood problem as follows: Let the likelihood function be given by:

$$L(\Theta) = P(D|\Theta) = \frac{\Gamma(N)}{\prod_{\rho \in \mathcal{M}} \Gamma(N\alpha_\rho)} \prod_{\rho \in \mathcal{M}} \Theta_\rho^{\alpha_\rho N}$$

MLE 1 *Find $\rho \in A$ so that $L(\rho) \geq L(\omega)\ \forall \omega \in A$.*

Similarly, for any specified set of loci $\{j_1, j_2, \dots, j_v\}$ we may define a likelihood function $L_{\mathcal{M}[j_1, j_2, ... j_v]}$ as the most likely to produce posterior $\alpha_{j_1, j_2, ..., j_v}$ over the space $\mathcal{M}[j_1, j_2, ...j_v]$.

Lemma 1 *If $d(\alpha, A) < \epsilon$ for some ϵ small enough, and $d(\alpha, A) = \min_{\theta \in A} ||\alpha - \theta||_2$. Maximizing $\prod_{\rho \in \{0,1\}^{M-1}} \Theta_\rho^{\alpha_\rho N}$ over $\Theta \in A$ is equivalent to minimizing*

$$\sum_{j \in [1...M]} \frac{(\alpha_j - \Theta_j)^2}{\alpha_j}$$

over $\Theta \in A$.

The proof in the appendix is derived from a Taylor-series expansion of the likelihood function. It demonstrates that the MLE result in set A is the vertex of a 2^{M-1} hyper-cube closest to our estimated joint probability function α, measured by a modified L_2 norm.

With this result we assume the following function to be used in the algorithms presented later:

Algorithm 1
MLE–COLLAPSE($j_1, j_2, \ldots, j_v$)
Compute $\rho \in A[j_1, j_2, \ldots, j_v]$
minimizing $\sum_{j \in [j_1, j_2, \ldots, j_v]} \frac{(\alpha_j - \Theta_j)^2}{\alpha_j}$ over $\Theta \in A$
return ρ

4 Algorithms

The algorithms focus on growing disjoint-phased contiguous sets of loci called *Contigs*. All loci are assigned an arbitrary phase and begin as a singleton phased contig. A JOIN operation checks if these phased contigs may be phased relative to one another using a function called VERIFY-PHASE. VERIFY-PHASE can be designed to check a phasing criteria, for example refuting a hypothesis of Hardy-Weinberg Equilibria is discussed in the appendix.

If a pair of phased contigs can be joined by passing the VERIFY-PHASE function then the disjoint sets are combined into a single phased contig and the joint distribution over the set is computed with the MLE–COLLAPSE function. Having completed a successful join operation, we may regard the distribution function as the most likely haplotypes generating the observed data over the specified loci. Because the growth of contigs is monotonic and depends on local information available at the time of the operation, in the full paper we also consider an ADJUST operation that fractures a contig and re-join's using a larger locality of data than what was available during the JOIN.

We describe the operations in detail, analyze the results and indicate how to avoid incorrect operations.

Collapse : In the previous section of this paper, we discussed the collapse operation as the MLE–COLLAPSE function. It may be used to update a joint probability distribution over a set of contigs; it has the effect of keeping the contig structures bound to haplotype states which simplifies the computing of a phase.

Join:

Let K be a parameter denoting neighborhood size. Letting $C_1 = \{j_1, j_2, \ldots, j_v\}$ and $C_2 = \{j'_1, j'_2, \ldots j'_w\}$, the join operation is as follows:

Given joint-probability functions $\alpha_{j_1}, \alpha_{j_2}, \ldots, \alpha_{j_v}, \alpha_{j'_1}, \alpha_{j'_2}, \ldots \alpha_{j'_w}$, compute the joint probability function α_{C_1,C_2} with formula

$$\alpha_{C_1,C_2} = \alpha_{(j_v)(j'_1 j'_2 \ldots j'_w)} = w_{j_v,j'_1}\alpha_{(j_v)(\bar{j}'_1 j'_2 \ldots j'_w)} + w_{j_v,j'_2}\alpha_{(j_v)(j'_1 \bar{j}'_2 \ldots j'_w)} + \ldots + w_{j_v,j'_K}\alpha_{(j_v)(j'_1 j'_2 \ldots \bar{j}'_K \ldots j'_w)}$$

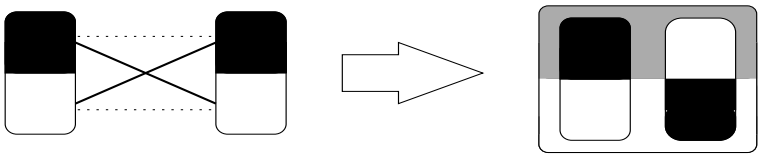

Fig. 1. Collapse

with

$$\alpha_{(j_v)(j'_1 j'_2 \ldots \bar{j}'_x \ldots j'_w)} = \sum_{i=1:N} \alpha^{(i)}_{j_v j'_x} = \sum_{i=1:N} \alpha^{(i)}_{j_v}(d_{ij_v})\alpha^{(i)}_{j'_x}(d_{ij'_x})$$

$$w_{j_v,j'_x} = \kappa \frac{1}{d(j_v, j'_x)}$$

Here, $\kappa = \frac{1}{\frac{1}{d(j_v,j'_1)}+\frac{1}{d(j_v,j'_2)}+\ldots+\frac{1}{d(j_v,j'_K)}}$ and $d(j_v, j'_x)$ is proportional to genomic distance between loci j_v and j'_x.

Algorithm 2
COMPUTE–PHASE($C_1 = \{j_1, j_2, \ldots, j_v\}$, $C_2 = \{j'_1, j'_2, \ldots, j'_w\}$, K)
assume j_v in C_1 is such that $d(j_v, C_2) \leq d(j, C_2)\ \forall j \in C_1$:
Compute $\alpha_{(j_v)(j'_1 j'_2 \ldots j'_w)}$ using parameter K.
return $\alpha_{(j_v)(j'_1 j'_2 \ldots j'_w)}$

Algorithm 3
JOIN($C_1 = \{j_1, j_2, \ldots, j_v\}$, $C_2 = \{j'_1, j'_2, \ldots, j'_w\}$, K)
COMPUTE–PHASE($C_1 = \{j_1, j_2, \ldots, j_v\}$, $C_2 = \{j'_1, j'_2, \ldots, j'_w\}$, K)
if (VERIFY–PHASE($\alpha_{j_1 j_2 \ldots j_v}$)) then
$\alpha_{j_1 j_2 \ldots j_v} \leftarrow$ MLE–COLLAPSE($j_1, j_2, \ldots, j_v$) ;

Our algorithm estimates the haplotypes by solving an ordered set of local MLE problems. The rationale of the chosen function is discussed in the full paper.

4.1 Implementation

`Input`

The input is a set of data points $\{d_{ij} \in \mathbb{R}^r : i \in [1 \ldots N], j \in [1 \ldots M]\}$. We make the following assumptions about the input:

– For each j the points $d_{1j}, d_{2j}, \ldots, d_{Nj}$ are derived from the Gaussian mixture model corresponding to mapping data at polymorphic loci j.

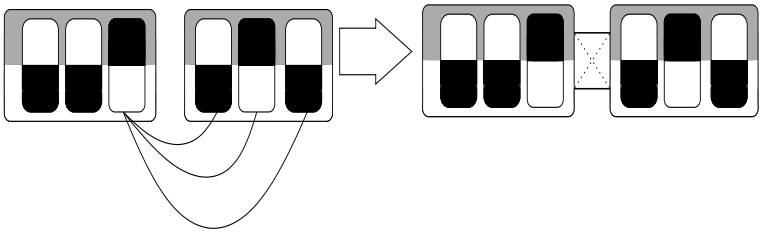

Fig. 2. Compute–Phase

– For each i points $d_{i1}, d_{i2}, \ldots, d_{iM}$ are independent random variables with parameters associated to underlying haplotypes.

Knowing the mapping order of polymorphic loci, we assume the positions of the genome to be $[x_1, x_2, \ldots, x_M]$.

`Pre-Process`

The EM-algorithm is run for each loci: $\{d_{ij} : i \in [1...N]$ observable $\} \rightarrow \{\hat{\Phi}_j : \alpha_j\} \quad \forall j \in [1, \ldots, M]$.

The result is a set of estimates for bi-allelic loci, $\{\hat{\Phi}_1, \hat{\Phi}_2, \ldots, \hat{\Phi}_M\}$, as well as a set of functions estimating the probability that any data point derives from the distinct alleles $\{\alpha_1, \alpha_2, \ldots, \alpha_M\}$.

Next we construct a join schedule. Letting $\beta_j = x_{j+1} - x_j$, we sort the results into an index array giving an increasing sequence: $\{j_1, j_2, \ldots j_v, \ldots j_{M-1}\}$.

`Main Algorithm and Data Structure`

Contigs are maintained in a modified union-find data structure designed to encode a collection of disjointed, unordered sets of loci which may be merged at any time. Union-find supports two operations, UNION and FIND [14]: union merges two sets into one larger set, FIND identifies the set containing a particular element. Loci j is represented by the estimated distribution α_j, and may reference its left and right neighbor. At any instant, a phased contig is represented by:

– A MLE distribution or haplotype assignment for the range of loci in the contig (if one can be evaluated).
– Boundary loci: Each contig has a reference to left- and right-most loci.

In the vth step of the algorithm, consider the set of loci determined by β_v, $\{j_v, j_{v+1}\}$: If FIND(j_v) and FIND(j_{v+1}) are in distinct contigs C_p and C_q, then 1) attempt to UNION C_p and C_q, by use of the JOIN operation and 2) update the MLE distribution and boundary loci at the top level if the JOIN is successful.

`Output`

Output is a disjointed collection of sets, each of which is a phased contig. It represents the most likely haplotypes over that particular region.

4.2 Time Complexity

The preprocess may involve using the EM–algorithm once for each loci. The convergence rate of the EM–algorithm is a topic of research ([8]) and depends on the amount of overlap in the mixture of distributions. For moderate-sized data sets we have noticed no difficulties with convergence of the EM-algorithm.

First we estimate the time complexity of the main algorithm implementing the $K-$neighbor version. For each β_{j_v} there are two find operations. The number of union operations cannot exceed the cardinality of the set $\{\beta_j : j \in [j_1, j_2, \ldots, j_{M-1}]\}$, as contigs grow monotonically. The time cost of a single find operation is at most $\gamma(M)$, where γ is the inverse of Ackermann's function. Hence the time cost of all union-find operations is at most $O(M\gamma(M))$. The join operation, on the other hand, requires running the $K-$neighbor optimization routine, at a cost of $O(K)$. Thus the main algorithm has a worst-case time complexity of

$$O\Big(M\left(\gamma(M)+K\right)\Big) = O\Big(M\gamma(M)\Big)$$

and may be regarded as almost linear in the number of markers, M for all practical purposes since K is almost invariably a small constant.

4.3 Examples

The appendix contains two examples illustrating the implementation for two simulated RFLP data sets, subject to extensive random errors.

5 Conclusions and Future Work

The simulation results are found to be encouraging, as they demonstrate that locally the phasing may be highly accurate. When local coverage derived from one haplotype is low, then the detection of polymorphisms become difficult. In the first data set a false negative detection is found on the 8th marker from the left, this is due to zero coverage from one of the haplotypes at that point. The ninth marker is a false negative detection and is attributed to zero coverage from one haplotype and low coverage (2 molecules) from the alternative haplotype. Note that the false positive does not cause errors in the phase information for correctly detected polymorphic loci in the phased–contig achieved over marker index in the set $\{7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$. Designing a mapping experiment targeting a polymorphic marker in the set $\{6, 7, 8, 9, 10\}$ could allow one to phase the two contigs into a single contig.

In the full paper we shall explore how the order of *local* maximum likelihood problem solutions relate to the *global* maximum likelihood problem. We further discuss mapping technologies and single molecule methods that may be used to generate data suited for the diploid haplotyping problem. In particular SNPs, micro–arrays, and DNA-PCR-Colonies or polonies ([9]) are under investigation by the authors. Further analysis and examples will be presented.

References

[1] T.S. Anantharaman, B. Mishra, and D.C. Schwartz. "Genomics via Optical Mapping II: Ordered Restriction Maps," **Journal of Computational Biology**, **4**(2):91–118, 1997.

[2] V. Bafna,D. Gusfield,G. Lancia, and S. Yooseph. "Haplotyping as Perfect Phylogeny, A Direct Approach," **Technical Report UC Davis** CSE–2002–21.

[3] W. Casey, B. Mishra, and M. Wigler. "Placing Probes on the Genome with Pairwise Distance Data," **Algorithms in Bioinformatics: first international workshop: proceedings WABI 2001**, 52–68, Springer, New York, 2001.

[4] A. Clark. "Inference of Haplotypes from PCR-Amplified Samples of Diploid Populations," **Mol. Biol. Evol**, **7**:111–122,1990.

[5] A. Dempster, N. N. Laird, and D.Rubin. "Maximum likelihood from incomplete data via the EM algorithm," **J.R. Stat. Soc.**, **39**:1–38, 1977.

[6] L. Excoffier, and M. Slatkin. "Maximum–Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population," **Mol. Biol. Evol**, **12**:921–927, 1995.

[7] D. Gusfield. "Inference of Haplotypes from Samples of Diploid Populations: Complexity and Algorithms," **Journal of Computational Biology**, **8**(3):305–323, 2001.

[8] J. Ma, L. Xu, and M. Jordan. "Asymptotic Convergence Rate of the EM–Algorithm for Gaussian Mixtures," **Neural Computation**, **12**(12):2881–2907, 2000.

[9] R. Mitra, and G. Church. "In situ localized amplification and contact replication of many individual DNA molecules," **Nucleic Acids Research**, **27**(24):e34–e34, 1999.

[10] T. Niu, Z. Qin, X. Xu, and J. Liu. "Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms," **Am. J. Hum. Genet. 70**:156–169, 2002.

[11] L. Parida, and B. Mishra. "Partitioning Single-Molecule Maps into Multiple Populations: Algorithms And Probabilistic Analysis," *Discrete Applied Mathematics*, (The Computational Molecular Biology Series), **104**(l-3):203–227, August, 2000.

[12] S. Roweis, and Z. Ghahramani. "A Unifying Review of Linear Gaussian Models," **Neural Computation**, **11(2)**:305–345, 1999.

[13] M. Stephens, N. Smith, and P. Donnelly. "A new statistical method for haplotype reconstruction from population data," **Am. J. Hum. Genet. 68**:978–989, 2001.

[14] R. E. Tarjan. **Data Structures and Network Algorithms**, CBMS 44, SIAM, Philadelphia, 1983.

[15] B. Weir. **Genetic Data Analysis II**, Sinauer Associates, Sunderland, Massachusetts, 1996.

Appendix

A RFLP Examples

We demonstrate our algorithm on two simulated data sets composed of ordered restriction fragment lengths subject to sizing error. Figure 3 below is presented in bands:

– The band nearest the bottom in the layout is the simulated haplotypes.

- The second band from the bottom is the haplotype molecule map for a diploid organism. These molecules (which are sorted into two haplotype classes in the layout) are mixed and made available to the algorithm as a single set of genotype data.
- The third band from the bottom shows the results of the EM-algorithm and the set of markers that are determined to have polymorphic alleles.
- The forth band in the layout provides the history of contig operations. From this tree one can view: 1) the developing k-neighborhoods, and 2) the distinct phased contigs.
- The top band in the layout gives the algorithmic output for this problem, including phased-in subsets that span the distance indicated by the bars above and below the loci markers. Areas where phase structure overlaps but cannot extend are regions that are of interest to target with more specific sequences, in order to extend the phasing.

Parameters of the simulations are summarized in the table:

Parameter	Symbol	Data Set 1	Data Set 2
Number of molecules	M	80	150
Number of fragments RFLP and non RFLP	F	20	100
Size of the genome	G	12000	50000
Expected molecule size	EMS	2000	2000
Variance in molecule size	VMS	50	500
Variance in fragment length size	VFS	1	20
P–value that any given fragment is an RFLP	P–BIMODE	.5	.3
Expected separation of means for RFLP	ERFLPSEP	10	50
Variance in the separation of means for RFLP	VRFLPSEP	.01	6

Any parameter with both an expectation and variance is generated with a normal distribution.

We used a simple Verify–Phase function which merely checked that our posteriori distribution α_{C_a,C_b} is separated by a distance of $C > 0$ from the point $[\frac{1}{2}, \frac{1}{2}]$. In practice we discovered that the parameter C should depend on the local coverage.

For the first simulation on data set I seen in figure A, a relatively small set is chosen so that the limitations of the algorithm can be seen. Here the neighborhood size is set to $k = 5$. There is no guarding against false positive RFLP detections, still phasings are computed and one can see that mistakes are due to the low coverage library.

In the second simulation on data set II seen in figure A we illustrate that good phasing results may be achieved on large, sparse data sets.

B MLE Estimate

If $d(\alpha, A) < \epsilon$ *for some* ϵ *small enough. Maximizing* $\prod_{\rho \in \{0,1\}^{M-1}} \Theta_\rho^{\alpha_\rho N}$ *over* $\Theta \in A$ *is equivalent to minimizing* $\sum_{j \in [1 \ldots M]} \frac{(\alpha_j - \Theta_j)^2}{\alpha_i}$ *over* $\Theta \in A$.

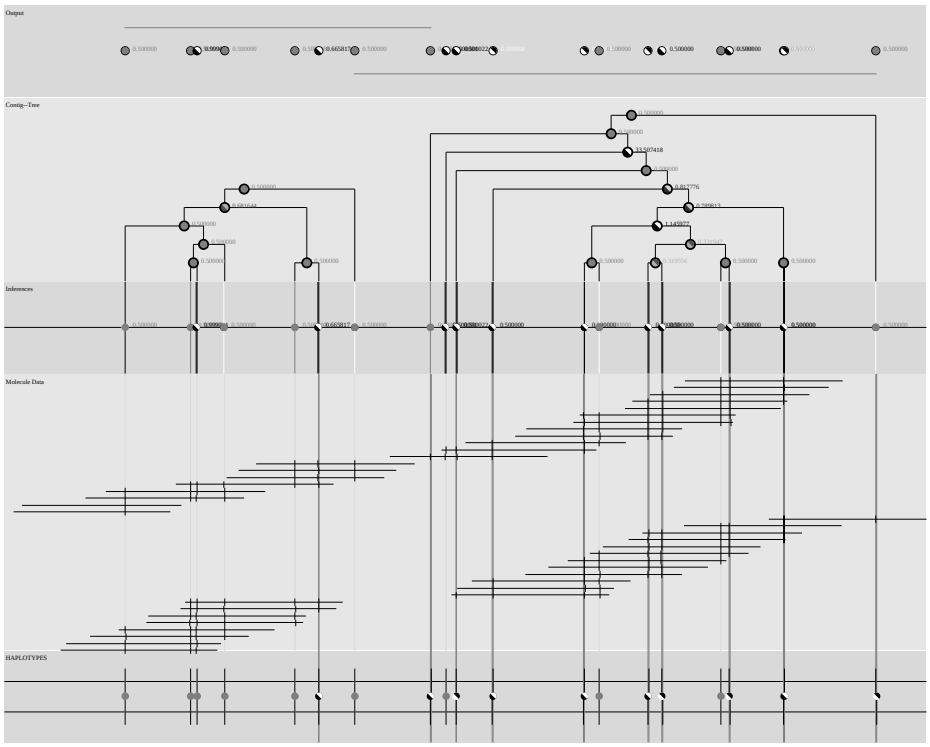

Fig. 3. Data set I

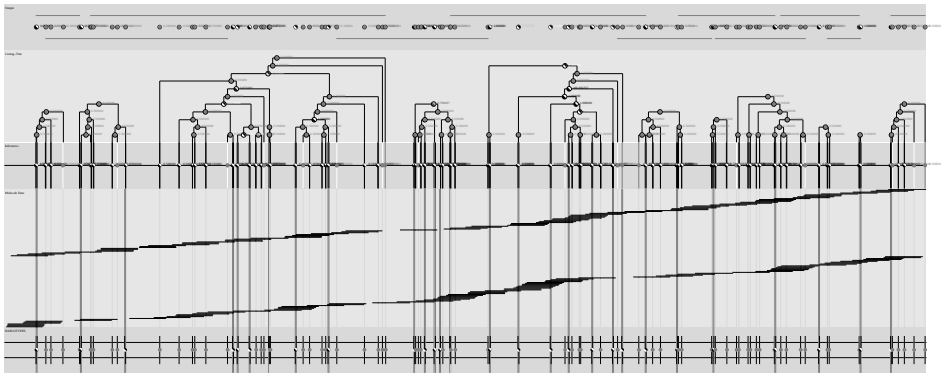

Fig. 4. Data set II

Proof. Let $F(\Theta) = \frac{n!}{\prod_{j=1:k} n_j!} \prod_{j=1:k} \Theta_j^{n_j}$. Computing the second variation:

$$F''(\theta) = \left(\begin{bmatrix} \frac{n_1^2}{\theta_1^2} & \frac{n_1 n_2}{\theta_1 \theta_2} & \dots & \frac{n_1 n_k}{\theta_1 \theta_k} \\ \frac{n_2 n_1}{\theta_2 \theta_1} & \frac{n_2^2}{\theta_2^2} & \dots & \frac{n_2 n_k}{\theta_2 \theta_k} \\ \vdots & \vdots & \dots & \vdots \\ \frac{n_k n_1}{\theta_k \theta_1} & \frac{n_k n_2}{\theta_k \theta_2} & \dots & \frac{n_k^2}{\theta_k^2} \end{bmatrix} - \begin{bmatrix} \frac{n_1}{\theta_1^2} & 0 & \dots & 0 \\ 0 & \frac{n_2}{\theta_2^2} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \frac{n_k}{\theta_k^2} \end{bmatrix} \right) F(\theta)$$

Since $F(\theta)$ is smooth in θ, Taylor's remainder theorem gives,

$$F(\Theta) = F(\alpha) + \nabla F(\alpha) \cdot (\Theta - \alpha) + (\Theta - \alpha)^T F''(\alpha)(\Theta - \alpha) + o\left(||(\Theta - \alpha)||_2\right)$$

When $\alpha = [\frac{n_1}{n}, \dots \frac{n_k}{n}]$, $\nabla F(\alpha) = 0$, this is a standard MLE result for a multinominal distribution. Computing the quadratic function:

$$\begin{aligned} &(\Theta - \alpha)^T F''(\alpha)(\Theta - \alpha) \\ &= (\Theta - \alpha)^T \left(n^2 \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} - n \begin{bmatrix} \frac{1}{\alpha_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\alpha_2} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \frac{1}{\alpha_k} \end{bmatrix} \right) F(\alpha)(\Theta - \alpha) \\ &= F(\alpha) n^2 \left(\sum_j (\Theta_j - \alpha_j) \right)^2 - F(\alpha) n \sum_j \frac{(\Theta_j - \alpha_j)^2}{\alpha_j} \\ &= -F(\alpha) n \sum_j \frac{(\Theta_j - \alpha_j)^2}{\alpha_j} \end{aligned}$$

Thus for an θ very near to α the level curves of F are given by $\Theta_\delta = \{\theta : F(\theta) = F(\alpha) - \delta\}$ are approximately ellipsoids.

$$\begin{aligned} F(\Theta) &= F(\alpha) - F(\alpha) n \sum_j \frac{(\Theta_j - \alpha_j)^2}{\alpha_j} + o\left(||(\Theta - \alpha)||_2\right) \\ &= F(\alpha) - F(\alpha) n \sum_j \frac{(\Theta_j - \alpha_j)^2}{\alpha_j} + o\left(\sum_j \frac{(\Theta_j - \alpha_j)^2}{\alpha_j} \right) \end{aligned}$$

Let $||\Theta - \alpha||_\alpha^2 = \sum_j \frac{(\Theta_j - \alpha_j)^2}{\alpha_j}$. Letting $L_1 = L(\alpha)$ and assuming there is a second local optima for the likelihood function value at L_2, let $V(\alpha) = \{\Theta : L(\Theta) > L_1 - \frac{L_1 - L_2}{2}\}$. We must show that there is a δ so that $\{\Theta : ||\Theta - \alpha||_\alpha^2 < \delta\} \subset V(\alpha)$. And this is clear from the inequality

$$L_1 - L_1 n ||\Theta - \alpha||_\alpha^2 - o\left(||\Theta - \alpha||_\alpha^2\right) < F(\Theta) < L_1 - L_1 n ||\Theta - \alpha||_\alpha^2 + o\left(||\Theta - \alpha||_\alpha^2\right)$$

by choosing a δ small enough that $L_1 n\delta + o(\delta) < \frac{L_1 - L_2}{2}$. We conclude that if there is a point of $A \in \{\Theta : ||\Theta - \alpha||_\alpha^2 < \delta\}$ then it must be the unique maxima in A for our likelihood function.

C Test for Hardy-Weinberg Equilibria at Different Loci

The Chi–Squared statistical test for determining whether allelic data at loci j and j' display linkage disequilibrium and hence are not in Hardy-Weinberg Equilibrium (HWE) have been very well studied. We refer the reader to Weir ([15]) for details and a complete statistical treatment.

We slightly modify the Chi-Squared statistical test for Gametic Disequilibrium at two loci using additive disequilibrium coefficients to adjust to our population model. The end result is a Chi-Squared statistical test that allows us to reject HWE from observed frequencies alone. Since determining linkage is a prerequisite to phasing, or at least in finding structure in the joint distribution over allele spaces of adjacent loci, the statistical test is important. The boundaries of haplotype blocks (or phased contigs, as we call them) are an interesting and important problem in understanding population dynamics.

Let D_{ab} denote the disequilibrium coefficient between alleles a at loci j and b at loci j':

$$D_{ab} = p_{ab} - p_a p_b$$

Where p_{ab}, p_a, p_b are the population frequencies for allele type: ab, a, b respectively. In the presence of HWE D_{ab} is expected to be zero. Letting $\hat{D}_{ab}$ denote an estimate from estimate frequencies:

$$\begin{aligned}\hat{D}_{ab} &= \tilde{p}_{ab} - \tilde{p}_a \tilde{p}_b \\ &= \frac{1}{N} \sum_{i=1:N} (\alpha_{ja}(d_{ij}) \alpha_{bj'}(d_{ij'})) - (\tilde{p}_a \tilde{p}_a)\end{aligned}$$

with:

$$\tilde{p}_a = \tfrac{1}{N} \textstyle\sum_{i=1:N} \alpha_{aj}(d_{ij}),\ \tilde{p}_b = \tfrac{1}{N} \textstyle\sum_{i=1:N} \alpha_{bj'}(d_{ij'})$$

Computing of Expectation and Variance:

$$\begin{aligned}E(\hat{D}_{ab}) &= \frac{N-1}{N} D_{ab} \\ V(\hat{D}_{ab}) &\approx \frac{1}{N} \left[p_a q_a p_b q_b + (1 - 2p_a)(1 - 2q_a) D_{ab} - {D_{ab}}^2 \right]\end{aligned}$$

The variance can be computed using Fisher's approximate variance formula. Under the assumption that loci j and j' are in HWE we have $D_{ab} = 0$ and:

$$\begin{aligned}E_{HWE}(\hat{D}_{ab}) &= 0 \\ V_{HWE}(\hat{D}_{ab}) &= \frac{1}{N} \left[p_a q_a p_b q_b \right]\end{aligned}$$

From this information, one may construct a Chi-Squared test to evaluate the hypothesis that alleles a and b at loci j and j' are acting as they would if they were in HWE.

$$\chi^2_{ab} = z^2 = \frac{N\hat{D_{ab}}^2}{\tilde{p}^A_a \tilde{q}^A_a \tilde{p}^B_b \tilde{q}^B_b}$$

We may reject the HWE hypothesis correctly 9 times in 10 by using a reference value of $z^2 > 2.71$, or we may reject HWE correctly 99 times in 100 using reference values $z^2 > 6.63$. If alleles are linked by a haplotype, this test may be used as the VERIFY-PHASE function mentioned previously in this text.

D EM–Algorithm Analytic Results

D.1 An Example Using RFLP Markers

The data at loci j refers to the observed distances between restriction sites j and $j+1$, as they are derived from two haplotypes H_1 and H_2 with underlying genome distances μ_1 and μ_2. The distribution of data points for loci j is given by:

$$f_j(x) = \frac{1}{\sqrt{2\pi\mu^2_{j1}}} \exp\left(\frac{-(x-\mu_{j1})^2}{2\mu^2_{j1}}\right) \alpha_{j1}(x) + \frac{1}{\sqrt{2\pi\mu^2_{j2}}} \exp\left(\frac{-(x-\mu_{j2})^2}{2\mu^2_{j2}}\right) \alpha_{j2}(x)$$

We make a simplifying assumption that $\sigma = \frac{1}{2}(\mu_{j1} + \mu_{j2})$ so that f_j may be closely approximated by:

$$F_j(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu_{j1})^2}{2\sigma^2}\right) \alpha_{j1}(x) + \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu_{j2})^2}{2\sigma^2}\right) \alpha_{j2}(x)$$

For loci j the set of points $\{d_{ij} : i \in [1, 2, \ldots N]\}$ is data. We infer the model parameters $\Phi = \{\sigma, \mu_1, \mu_2\}$ and posteriori distribution α by use of the EM–algorithm. We drop the subscript j in the following equation, the objective being to iteratively optimize the function:

$$H(\alpha, \Phi) = \sum_{i \in 1:n} \sum_{k \in 1:2} \left(\alpha_k(d_{ji}) \ln G_k(d_{ji}|\Phi) - \alpha_k(d_{ji}) \ln\left(\alpha_k(d_{ji})\right)\right)$$

With $G_k(x|\Phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu_k)^2}{2\sigma^2}\right)$ the kth Gaussian kernel.

Optimization is done in two steps:

1. E–STEP Holding Φ fixed, optimize $H(\alpha, \Phi)$ over α, letting $\hat{\Phi}$ be the previous estimate of parameters.

The result for the argmax α is:

$$\alpha_k(x) \leftarrow \frac{G_k(x|\hat{\Phi})}{\sum_l G_l(x|\hat{\Phi})}$$

2. M–STEP

Holding α fixed, optimize $H(\alpha, \Phi)$ over Φ (using the previous estimate of parameters on the hidden categories $\hat{\alpha}(y)$, which depend on the previous estimate denoted $\hat{\Phi} = [\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}]$). The result $\operatorname{argmin} H(\hat{\alpha}, \Phi)$ is:

$$\mu_k \leftarrow \frac{\sum_{i=1:n} \hat{\alpha}_k(d_{ij}) d_{ij}}{\sum_{i=1:n} \hat{\alpha}_k(d_{ij})}$$
$$\sigma \leftarrow \sqrt{\frac{1}{2} \sum_{j=1:2} \frac{1}{N} \sum_{i=1:N} \hat{\alpha}_j(d_{ij})(d_{ij} - \hat{\mu}_j)^2}$$

The EM-algorithm is run until convergence in the parameter space occurs. Detailed computations for the E-Step and M-Step are provided in the appendix.

Detailed proofs of each step are found below.

E-Step

Proof. Consider the calculus problem of optimizing:

$$f(\phi) = \phi\,(A_1 - \log(B\phi)) + (1-\phi)\,(A_2 - \log(B(1-\phi)))$$
$$f'(\phi) = 0 \Rightarrow \phi^* = \left(\frac{1}{e^{A_1 - A_2} + 1}\right)$$

Notice that $\phi^* \in (0,1)$. Apply this fact to the optimization problem of finding numbers $\hat{Q} = \langle \alpha_{i\nu} \rangle_{i=1:N, \nu=1:2}$, so that the following function is optimized:

$$\sum_{i=1:n} \sum_{\nu=1:2} (\alpha_{i\nu}\,(A_{i\nu} - \log(B\alpha_{i\nu})))$$
$$= \sum_{i=1:n} \alpha_{i\nu}\,(A_{i1} - \log(B\alpha_{i\nu})) + (1-\alpha_{i\nu})\,(A_{i2} - \log(B(1-\alpha_{i\nu})))$$

Where $A_1 = \frac{(d_i - \mu_1)^2}{2\sigma^2}$ and $A_2 = \frac{(d_i - \mu_2)^2}{2\sigma^2}$ and $B = \sqrt{2\pi\sigma^2}$. We see that the answer is given by maximizing each summand and hence given by:

$$\alpha_1 = \left(\frac{1}{e^{\left(\frac{(d_i-\mu_1)^2}{2\sigma^2} - \frac{(d_i-\mu_2)^2}{2\sigma^2}\right)} + 1}\right)$$
$$= \frac{G_1(d_i)}{G_1(d_i) + G_2(d_i)}$$

and similarly for α_2.

M-Step

Proof. Consider the calculus problem of optimizing:

$$\begin{aligned} f(\mu_1, \mu_2, \sigma) &= \sum_{i=1:N} \sum_{\nu=1:2} q_{\nu i} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(a_i - \mu_\nu)^2}{2\sigma^2} \right) + H \\ &= \sum_{i=1:N} \sum_{\nu=1:2} q_{\nu i} \left(\frac{-(a_i - \mu_\nu)^2}{2\sigma^2} - \log \sqrt{2\pi\sigma^2} \right) + H \end{aligned}$$

Where H is constant in (μ_1, μ_2, σ). Consider the partial derivative of f with respect to μ_ν:

$$\begin{aligned} \frac{\partial f}{\partial \mu_\nu} &= \sum_{i=1:N} \frac{\partial}{\partial \mu_\nu} \Big(\frac{q_{\nu i}}{2\sigma^2} (-a_i^2 + 2a_i\mu_\nu - \mu_\nu^2 - 2\sigma^2 \log \sqrt{2\pi\sigma^2}) \Big) \\ &= \frac{1}{\sigma^2} \sum_{i=1:N} q_{\nu i} a_i - \mu_\nu q_{\nu i} \end{aligned}$$

Thus we get:

$$\frac{\partial f}{\partial \mu_\nu} = 0 \Leftarrow \mu_\nu = \frac{\sum_{i=1:N} q_{i\nu} a_i}{\sum_{i=1:N} q_{i\nu}}$$

Now consider the partial of f with respect to σ:

$$\begin{aligned} \frac{\partial f}{\partial \sigma} &= \sum_{i=1:N} \sum_{\nu=1:2} q_{\nu i} \left(\frac{(a_i - \mu_\nu)^2}{2\sigma^2} - 1 \right) \\ &= \frac{1}{2\sigma^2} \sum_{i=1:N} \sum_{\nu=1:2} q_{\nu i} \left((a_i - \mu_\nu)^2 - 2\sigma^2 \right) \end{aligned}$$

Thus we get:

$$\frac{\partial f}{\partial \sigma} = 0 \Leftarrow \sigma = \sqrt{\frac{1}{2} \sum_{\nu=1:2} \frac{1}{N} \sum_{i=1:N} q_{i\nu} (a_i - \mu_\nu)^2}$$

Hence the necessary condition for $(\mu_1^*, \mu_2^*, \sigma^*)$ to be the maximizing argument is that:

$$\begin{aligned} \mu_1^* &= \frac{\sum_{i=1:N} q_{i1} a_i}{\sum_{i=1:N} q_{i1}} \\ \mu_2^* &= \frac{\sum_{i=1:N} q_{i2} a_i}{\sum_{i=1:N} q_{i2}} \\ \sigma^* &= \sqrt{\frac{1}{2} \sum_{\nu=1:2} \frac{1}{N} \sum_{i=1:N} q_{\nu i} (a_i - \mu_\nu)^2} \end{aligned}$$